

## DOCUMENT RESUME

ED 194 553

TM 800 682

AUTHOR Thorndike, Robert L.  
TITLE The Application of Bayesian Thinking to Educational Measurement Problems.  
PUB DATE 80  
NOTE 9p.  
EDRS PRICE MF01/PC01 Plus Postage.  
DESCRIPTORS \*Bayesian Statistics; \*Computer Assisted Testing; Error of Measurement; Foreign Countries; \*Probability; \*Testing; Test Theory  
IDENTIFIERS \*Adaptive Testing; Australia; Stradaptive Tests

## ABSTRACT

In an invitational address to the Victorian Institute of Educational Research, the author discussed Bayesian theory and its relationship to the design and construction of tailored or adaptive tests. Bayesian thinking involves recognizing the role of prior probabilities and using these probabilities in combination with new data to arrive at future probabilities. These prior probabilities should be viewed as beliefs held in common agreement by individuals, based on data which has been carefully gathered, thoughtfully analyzed, and widely disseminated. Bayesian statistics quantify the prior probabilities, and these probabilities combined with subsequent data, such as test scores, can be used to arrive at a valid posterior judgment. By using Bayesian concepts and computers, tailored tests can be geared to individual examinees' estimated ability levels. Even without computer-assisted testing and large item banks, testing can be made more efficient by: (1) using a multi-level test format; (2) allowing the examinee to select the difficulty level of the test; or (3) employing a two-stage test, in which a routing test is given first to determine the difficulty level of the main test. Stradaptive testing was a method once considered, but was abandoned as too difficult to administer. (MH)

\*\*\*\*\*  
\* Reproductions supplied by EDRS are the best that can be made \*  
\* from the original document. \*  
\*\*\*\*\*

ED194553

TM800682

## The Application of Bayesian Thinking to Educational Measurement Problems

Robert L. Thorndike

Professor Thorndike was invited by ACER to give the keynote address at the Invitational Seminar on the Improvement of Measurement in Education and Psychology held in May 1980 to mark the 50th Anniversary of the founding of ACER. While in Melbourne, Professor Thorndike presented the John Smyth Memorial Lecture, sponsored annually by the Victorian Institute of Educational Research. The lecture is reproduced with the permission of the Institute.

U.S. DEPARTMENT OF HEALTH,  
EDUCATION & WELFARE  
NATIONAL INSTITUTE OF  
EDUCATION  
THIS DOCUMENT HAS BEEN REPRO-  
DUCED EXACTLY AS RECEIVED FROM  
THE PERSON OR ORGANIZATION ORIGIN-  
ATING IT. POINTS OF VIEW OR OPINIONS  
STATED DO NOT NECESSARILY REPRESENT  
OFFICIAL NATIONAL INSTITUTE OF  
EDUCATION POSITION OR POLICY.

It was some 200 years ago, plus a bit, about the time that Captain Cook was charting the shores of your fine country, that the Reverend Thomas Bayes set forth the theorem that has subsequently borne his name. That he had some reservations about his theorem is suggested by the fact that he did not publish a report of it during his lifetime. It was only after his death that his paper was read by a friend before the Royal Society of London, in 1763 and this was perhaps prophetic because Bayes' Theorem, though accorded an honorable place in the formal study of probability theory, was viewed with suspicion and considered somehow disreputable by statisticians and by scientists, and had little impact upon their thinking until quite recently. In the past 15 or 20 years, however, a new interest in Bayesian approaches has developed, and especially in educational research and in educational and psychological measurement Bayesian thinking appears to be assuming an important role.

This evening I would like to try to introduce you to the Bayesian type of thinking, which tends to be very closely akin to a common sense type of thinking. Then I would like to indicate how this sort of thinking applies to certain situations involving the interpretation of test data and research results. Finally, I would like to show how this approach relates to the design and construction of what are called 'tailored' or 'adaptive' tests that permit gathering information about each individual's abilities in a fraction of the time required by conventional testing instruments.

To develop some of the notions underlying Rev. Bayes' Theorem, let us set forth an imaginary (but I think not unrealistic) example. Let us suppose that we have, for a period of years, been administering to all students entering our college a self-descriptive personality inventory that yields a number of scores indicative of potential personal difficulties, one of these scores being an indicator of susceptibility to feelings of depression and anxiety. Let us suppose that over the years the college health office has kept a record of all the instances of actual or attempted or threatened suicide. Let us suppose that half of what for conciseness sake we shall call suicide cases showed, as entering freshman, a very high score on the depression scale of our inventory - a score so high that it occurred in only 1 in 20 of the total entering group. It would appear that we have a very good diagnostic indicator for picking in advance those students who will attempt suicide - 50% high scores among the suicides, as opposed to only 5% in the total entering group.

So now we ask the pay-off question: Knowing that a person shows this diagnostic sign - a high depression score - what can we say about the probability that that person will, sometime during his or her college career, attempt suicide? This question is sometimes labelled the problem of inverse probability. We know the likelihood of the symptom (high depression), given the event (suicide). We seek to know the probability of the event (suicide), given the symptom (high depression score).

Unfortunately, as it stands, the question is unanswerable. It can only be answered when we know the prior probability of the event in the total population, in this case in the total group of entering students. What is the overall rate of suicide, attempted suicide and threatened suicide in our total group of college students? A Bayesian would speak of this as the prior probability, a medical statistician as the base rate.

Let us suppose that overall the rate of suicides, attempted suicides or threatened suicides is one per hundred students. Then Bayes' Theorem provides the answer to our question about the probability of suicide, given a high depression score on our inventory. The posterior probability, that is, the probability, given that the indicator is present, is equal to the prior probability times the likelihood of the indicator, given the event, divided by the overall likelihood of the indicator. In our illustration it is .01, the prior probability, times .50, the likelihood of high depression score among suicides, divided by .05, the overall likelihood of high depression scores. The end result of the multiplications and divisions is to yield a posterior probability of .10, that is, one chance in 10 that the person with the very high depression score will in fact attempt suicide. By all odds, our best bet, even given the high depression score, is that he will not.

What this illustration points out is that your final betting odds on any event are dependent not only on the information available to you about the specific case, but also upon the prior odds on that event. When the event is an extremely unlikely one, even a strong diagnostic sign may not make its occurrence a good bet. It is this explicit recognition of the role of prior probabilities, and insistence upon being clear and forthright about prior probabilities that lies at the heart of Bayesian thinking. Procedures for clarifying one's prior probabilities and rules for combining prior probabilities with new data to arrive at a final statement of the posterior probabilities constitute the essence of Bayesian statistical method.

Also central to Bayesian thinking is the view that probabilities are to be thought of as beliefs held by individuals, rather than absolutes existing apart from the individuals who entertain them. On problems where data have been carefully gathered, thoughtfully analysed and widely disseminated, these beliefs will be in substantial agreement from one person to another. When little evidence is available, persons may show substantial variation in their beliefs about a given phenomenon. The role of research, to the Bayesian, is to provide a body of evidence that will correct discordant prior beliefs and bring persons closer and closer together in the posterior beliefs that they hold.

We do have prior beliefs about a wide range of phenomena in this world, held with varying degrees of conviction, and based in varying degrees upon evidence. Thus, if I asked each one of you to estimate what the maximum temperature would be at the Melbourne weather station tomorrow, each of you would feel fairly comfortable in giving me a value, your estimates would be in fairly good agreement, and none of them would be absurd - like -10 Celsius. You might very well protest that you couldn't specify an exact figure for tomorrow's temperature, but you could give a range within which the temperature would probably fall - and this is characteristic of many of our prior beliefs. They are well represented not by a point value, but by a range of possible values, with some being viewed as more likely than others.

Let's see how we might quantify an expression of your prior probabilities. A Bayesian would be inclined to express them in terms of betting odds - a concept that I have been led to believe is not totally unfamiliar in your country. If we want you to be thoughtful about your probabilities on tomorrow's temperature, let us say that you will be permitted to pick a value, you and I will each put up \$5.00, and I will have first choice as to whether I want to bet on the thermometer going above or staying below the figure that you pick. (If it falls exactly at the mark, the bet is off.) You will have to take the side of the bet that is left.

(Some of you who have had two children and only one piece of cake may recognize here the familiar parental trick. 'Mary, you can cut the cake, but John gets first pick'.) Under these circumstances, if you don't want to gradually go broke, you will set a value such that you think that higher and lower values are equally likely, so that my first choice can give me no advantage. The value that you pick becomes a good operational definition of the median or mid-point of your prior probability distribution. Extending the same kind of thinking, I can specify that you are to state a range of temperatures

within which you feel 95% sure that tomorrow's maximum will fall. In betting terms, one of us will put up \$1.00 and the other will put up \$19.00. You will specify the range of temperatures, but I will have the choice of whether to take the short odds or the long odds bet. If you set too wide an interval, I will put up \$19.00, take the bet that the temperature will fall within your range, and almost always win your dollar. If you set too narrow an interval, I'll put up a dollar on the bet that the temperature will fall outside your range, and win your \$19.00 often enough to show a profit.

The range that you set would be called by a Bayesian your '95% credible interval'. You would want to estimate it with some care before you staked much money on it.

We are not too familiar with the notion of formalizing our prior probabilities, and one of the things that Bayesians do is to try to get us to be more conscious of our prior probabilities, and to give us practice in quantifying them. Thus, a teacher who has worked with a class of 8-year-olds for several months will have some fairly definite ideas as to how well each of the children is likely to do on a standardized reading achievement test. Observing the children's class work and knowing how each has done on any informal tests and exercises, the teacher will have fairly definite ideas as to which are good readers, which are average readers, and which are minimally literate. If pushed, the teacher could probably produce for each child a median estimate, where higher or lower scores seemed equally likely, and even a 'credible interval' within which he or she was 95% certain that the child will fall. A Bayesian would recognize that the teacher's knowledge of his or her pupils was a significant datum, and would try to help the teacher to express that knowledge in clear and definite terms. The teacher might be encouraged to think of reading performance in terms of grade levels on a reading test and to produce for each child a grade equivalent level that would represent an even bet of the child falling above or below that level. The teacher might be coaxed to estimate a 95% credible interval - one where there was only about one chance in 20 that the child's test grade equivalent would fall outside that range.

Now the test is given and a score is obtained for each child. Often the test results come close to the teacher's prior median estimate; for some few children they do not. How should the teacher's or the school's posterior estimate of each child's ability put together the prior estimate from the teacher with the results from the test score? The teacher is not infallible, but neither is the test. A test is only a limited sample of each child's reading performance and we know that if a different form of the test had been given, or the test had been given at a different time, scores for specific children would have been somewhat different. Test publishers express that expected amount of variation as the standard error of measurement for the test. Given that standard error of measurement, we can establish a 95% credible interval for the score, just as we can for the teacher's estimate. Both the teacher's estimate and the test score are relevant evidence. A good Bayesian, whether teacher or administrator, would seek to combine the two sorts of evidence in a rational way so as to arrive at the soundest possible posterior judgment of each pupil's reading ability.

I shall not go into details of the procedure for combining prior estimates like the teacher's with subsequent data like the test score, but will simply state that, given that the teacher's estimates were carefully made, the most appropriate posterior point estimate for a pupil will be a weighted average of the teacher's prior estimate and the test score, the weights being in proportion to the narrowness of the credible interval for each. We can also determine an appropriate credible interval for the posterior estimate, and, not surprisingly, this interval will be narrower than either the credible interval of the teacher's judgment or the credible interval from the test. That is, by combining the two types of evidence, we arrive at a conclusion that can be viewed as more precise than a conclusion permitted by either of them taken separately.

A somewhat similar opportunity for Bayesian thinking arises when our prior estimate is based not on a teacher's judgment but on some earlier test result. Suppose we are now

about to give a reading test in the 4th grade to a group of pupils who have received a similar test in the 3rd grade. There is, naturally, a substantial correlation between 3rd grade and 4th grade test performance. Perhaps we know how close this relationship is from information in the technical manual that accompanies the test; perhaps we know it from earlier longitudinal studies in our own school. From whichever source, we have set up the equation that yields an estimate of 4th grade performance from earlier performance in the 3rd grade, and we have calculated the standard error of that estimate. We also know the standard error of measurement for the 4th grade test. Now we have Henry, whose predicted 4th grade score on the basis of his 3rd grade test result is a grade equivalent of 4.5. To our pleased surprise, when actually tested in the 4th grade, he obtains a grade equivalent of 5.4. What is our best estimate of his true ability at the time of the 4th grade test? We should not ignore the prior evidence from the 3rd grade, but how much weight should we give the prior evidence relative to the weight for the current test?

The Bayesian key to this situation lies in a comparison of the two standard errors - of estimate from the 3rd grade test and of measurement on the 4th grade test. The larger the standard error (and consequently the credible interval), the less the weight that should be given to that datum. It turns out that the weights should be in inverse proportion to the squares of the standard errors. Thus, if the standard error of estimate from grade 3 is 7/10 of a grade equivalent, and the standard error of measurement in grade 4 is 5/10 of a grade equivalent, it works out that the 4th grade results should receive twice as much weight as the year-old 3rd grade results, and our best estimate of Henry's true reading ability is a grade equivalent of  $(4.5 + 2$

$(5.4))/3$  or 5.1. The standard error of this estimate is equal to  $\sqrt{\frac{1}{\frac{1}{0.7^2} + \frac{1}{0.5^2}}}$

which equals 0.41. The gain in precision in this example, with the figures as given, is fairly modest, but the precision resulting from using both the earlier and the later tests, with weights appropriate to each, is clearly greater than that possible from either taken separately. It does pay to give heed to prior probabilities.

Let me give one further illustration of Bayesian thinking in a slightly different context before turning to procedures for adaptive testing. Suppose that we have introduced a new somewhat specialized program in our school, perhaps a course in computer programming, and we would like to know how effective the Handy Dandy Programmer Aptitude Test is for identifying students who will do well in that course. So we arrange to give the Handy Dandy at the beginning of the year to all who enroll for the course, lock up the results, and eventually correlate them with final grade in the programming course. But we have only 12 students in the course this year! The measure of correlation based on a sample as small as this is woefully unstable. What to do?

One thing we could do, of course, would be to continue to gather data for 5 or 10 years, in order to build up the size of our sample. As good Bayesians, we would start with a prior estimate of what we thought the correlation would be likely to be, and would modify that each year with the addition of new data. Eventually we would have accumulated enough data so that our credible interval was down to acceptably narrow limits, and then we would express our judgment of how satisfactory the Handy Dandy was as a device to select students for the programming course - if we hadn't died or changed the course first.

But another alternative might be to locate some other schools with similar courses that had used or were currently using the Handy Dandy, and to get reports of the correlation of test with grades for their students. We might not be sure that the course or the students in those schools were exactly the same as ours, so we might decide to weight each student from other schools rather less than one of our own students, but if we could quickly share results with a dozen other schools where there were, in all, 150 students we might decide that the 150 students should receive as much weight as we

would give to 75 of our own. Combining a prior estimate based on the average correlation in those schools with whatever result we found for our own 12 students would permit us immediately to arrive at a posterior estimate with precision equivalent to that from 87 students - a result that it would have taken us seven years to reach if we had relied solely upon our own resources. There are all sorts of ways that we can refine our prior estimate, before turning to our own, often limited data.

Let me turn now to adaptive testing, show how and why it works, and how it fits into the general model of Bayesian thinking. It seems fairly obvious that we test a person in order to obtain additional information about that person. We value a test in proportion as it tells us something that we don't already know. We don't ask a secondary school graduate: 'How much is  $2 + 2$ ?' Partly we don't because we would look silly, but mostly we don't because the answer 'Four' would tell us nothing about the person that we didn't already know (excepting that the person is tolerant of our foolishness and co-operative). Any other response such as 'five stupid!' would tell us more about the respondent's personality than about his cognitive processes. By the same token, we don't ask a 6-year-old: 'What is the cube root of 27?' we already know that the chance is infinitesimal that we will get the correct answer. For useful testing we ask questions at a level where we don't know in advance whether the examinee will be able to respond correctly. And the more uncertain we are initially, the more we learn about the person from the fact that he does (or does not) answer the question correctly. The maximum uncertainty exists when the prior probability of answering the question is exactly 50-50, so that we would be equally willing to bet on either side of the issue. (As a slight aside, the 50-50 applies if the response is generated by the examinee. Where the examinee selects from two or more provided alternatives, the optimal percentage depends upon the number of answer choices provided.)

Test theoreticians have, in the past 10 or 20 years, started to talk about the information function of a test item. This concept relates to the amount of information that the item provides us for a person at some specified level of ability. The information from an item depends on the one hand upon how close the item is to the 50-50 split for a person at that level of ability, and, on the other, on how steeply the probability of success rises as we go up the ability scale. No item is effective in providing information at all levels of ability. Unfortunately, the item that is most informative for persons of low ability is likely to be least informative for those of high ability, and vice versa. In the one case, we know in advance that persons of this ability level are almost certain to get the item right so we learn nothing when they pass, and in the other case that they are almost certain to get the item wrong, so we learn nothing when they fail.

A test that is designed to be used with all the pupils in a given grade, or perhaps even over several grades, is almost necessarily quite inefficient. If the test is to provide some information at all levels, a number of the items must be so easy that they tell us little or nothing we don't already know about the ablest students; a number must be so difficult that they tell us little or nothing about the least able. What we would like to be able to do is to test each examinee with those test exercises that would provide the most information about her or him.

If we can cross-breed our Bayesian concepts with a high speed computer, this tailoring of the test to the individual examinee becomes quite possible. We store in the computer's memory a data bank of test items whose difficulty level and whose effectiveness in differentiating among levels of ability is known from previous try-out, and we program the computer to control the order in which test exercises from this pool are presented to the examinee. Choice of the first exercise is based on our prior estimate of the examinee's ability. In the simplest case, when we have no specific advance information about each examinee, we may use the same starting place for all examinees in a given category - for instance, all 10-year-olds, all army recruits, or all applicants for a job as bank teller. The first item is displayed on a video screen and the

examinee responds. If the response is correct, the computer notes this and makes an immediate adjustment in its estimate of the person's ability level. (The size of the adjustment depends upon how much prior information the computer has, and how much information is provided by the present test item.) We can think of the adjusted value as a posterior ability estimate, with a credible interval for that examinee that has been slightly narrowed because of the added information from the item that has been answered.

The computer now searches its memory for the item that will provide the most information, given the present estimate of the examinee's ability level - a more difficult item if the first item was passed, an easier one if the first item was failed, and presents this item. The examinee responds to this second item, and at that point the cycle is repeated, updating the estimate of examinee ability, reducing the credible interval, and selecting another item. Very quickly, the computer's estimate is adjusted upward or downward, until the examinee is getting about half of the items right, and the adjustments are oscillating up and down in progressively smaller amounts.

Testing continues, at each stage picking the most informative item in the light of the computer's present estimate of the examinee's ability, until the credible interval has been reduced to an acceptably narrow range, or until some specified maximum number of items has been given. Experience with such computer-based testing indicates that a specified level of precision in the measurement of each examinee can usually be achieved with no more than half as many items as would be required in a conventional test. No time is wasted in giving items that provide little or no information about the particular examinee who is now face to face with the computer's terminal. Since the computer has adjusted its estimate of the examinee's ability after each test item, the final adjusted estimate constitutes the examinee's score on the ability measured by the test.

You will observe in this testing procedure all the elements of the Bayesian paradigm. We start with a prior estimate, if only a rough one that assigns to each examinee the prior distribution for a complete age group. We gather new data - in this case response to a test exercise with known properties. We combine our prior estimate with the data and get a posterior estimate with a certain mid-value and credible interval. We use our posterior from the first cycle as our prior for the next, and gather further data, at each stage choosing the datum that will give us the most information. We continue until the credible interval - it would be called a confidence interval in conventional statistics - is narrow enough to meet our specifications, and then we report our final posterior estimate.

'But', you say, 'I don't have a computer, and I don't have a huge bank of items from which to select as I test a particular examinee. What can I do to make my testing more efficient?' Half a dozen approaches have been proposed. I will identify and develop two or three that seem to me to be the most promising as practical and as approaching the efficiency of the exact Bayesian procedure.

Perhaps the simplest gesture towards adaptive testing is to develop a test in what we have called a 'multi-level' format. In this format, items of a given type, such as verbal analogy items, are arranged in a number of modules, each module of several items somewhat more difficult than the one before it. In our Cognitive Ability Tests, there are 12 modules in a given sub-test, but each examinee completes only five of them. The difficulty level of the test is controlled by the point at which the examinee starts and stops. The primary purpose of the multi-level format is to provide a graded series of tasks that are appropriate over a considerable range of ages or grades - in this instance, from eight-year-old to end of secondary school. However, if some individual is believed to be above or below average for his age or grade placement, it is a simple matter to administer a higher or lower level of the test than the standard one for that age or grade. Thus, one can use one's prior estimate of the child's ability level to assign a test level whose average difficulty will approximately match the child's ability.

It is important to recognize that in any of these forms of individualized testing the final scale value assigned to a person depends not simply upon the number of items right, but also on the difficulty of the items attempted. Thus, on our multi-level tests raw scores on each level are converted to a common 'universal scale score' and the expectation is that, though the raw score of number of items right would differ, a given individual would get essentially the same scale score no matter which level of the test was taken. The advantage of taking the most appropriate level is that the information will be maximized for that individual, and the measurement error reduced to a minimum.

A second procedure proposed for adapting test difficulty to individual ability is to let the examinee select for himself the difficulty of test exercises to be attempted. With a test in modular format, all examinees could be started at a module of moderate difficulty. If a given examinee felt the initial module was too difficult, she or he could move down to an easier module; if too easy, up to a more difficult one. Examinees would be encouraged to attempt modules that seemed to present a challenge, but not to be overpowering. The hope is that examinees would select levels on which they experienced a good mixture of success and failure, and which consequently provided a high level of information about them.

Such a testing procedure would seem to require a fairly high level of sophistication on the part of examinees if it were to work, and consequently to be limited to secondary school and above. Some limited trials of the procedure have suggested that examinees tend to choose items too easy to provide the maximum amount of information about them, and one suspects that this tendency will vary widely from one examinee to another, so that some will be quite poorly measured.

One approach that seems quite promising is two-stage testing. In this procedure, a brief routing test is given first, in order to determine the level of the main test that seems most appropriate for a given examinee. To be effective, the routing test must be one that will have a substantial correlation with the main test. Then prior probabilities are established, based on score on this short routing test, plus, perhaps, chronological age or any other facts about the examinee, to locate and narrow down the range of the prior credible interval. This information is used to define the level (or levels) of test items to be given in the main test. To provide a concrete example, let me describe how I see this two-stage procedure operating in an individual test that we are now working on as a potential successor to the long-established Stanford-Binet.

As a routing test, we shall probably use a steeply graded vocabulary test. This will start at the two-year level with a few pictures to be named, will move through simple words to be defined such as 'orange' and 'garden', and will continue up to words that few US college entrants can define, words such as 'dogmatic' and 'testator' - which most college freshmen think of as a person who gives tests. The complete routing test will consist of perhaps 30 items, and be expected to take something like five minutes. The examiner will have to make an immediate judgment on whether each response is acceptable and will stop after an examinee has failed two (or perhaps three) items. The examiner will then note the number of the last item attempted, and with that number and the examinee's age, will enter a table that will tell at which level to start the main test.

The main test will be arranged in about 20 levels, equally spaced in difficulty, covering the range from easy for two-year-olds to difficult for secondary school graduates. Items in a given level will be as nearly uniform in difficulty as our bank of item data permits, and each level will be clearly more difficult than the level below it. We expect that each examinee will be tested with three adjacent levels of the test. The initial level selected will be one on which the examinee is predicted to succeed on about 75 per cent of the items, on the next higher level the anticipated rate of success is about 50 per cent, and on the third level about 30 per cent. Thus, the tasks will be ones on which there is close to the maximum amount of prior uncertainty as to whether that examinee will get them right, and consequently ones that will provide a good deal of information with which to sharpen our posterior estimate of the individual's ability level.

This approach is not fool-proof, but we hope that school psychologists and others with similar levels of training will be able to handle the mechanics of the testing. The main goal is to compress the necessary testing time so that one can assess several distinct aspects of ability in the amount of testing time that would have been needed for only one.

One testing pattern that we considered for our work, but abandoned as probably too difficult for an examiner to handle, is what has been labelled a 'stradaptive' test. Here, as in our present plan, items are grouped into levels or strata on the basis of their difficulty. Each examinee is started at the stratum that corresponds roughly to the prior estimate of that examinee's ability. If the first item is passed, the examinee is moved up to an item at the next higher level of difficulty; if the item is failed, the examinee is moved down. Each time an item is passed, the examinee moves up, and each time one is failed he drops back. Generally, after the first few items, the examinee can be expected to oscillate up and down among the two or three levels that just about match his or her ability level, and each examinee will end up with close to 50 per cent of successes. The scale value assigned to the examinee will be determined primarily by the levels at which he or she is alternately passing and failing items.

This model seems about the closest approximation to the full Bayesian model of adjusting one's prior estimate for the examinee after each item. But it requires an immediate judgment of the correctness of each response. It also requires fairly complex bookkeeping to keep track of which items have been given and which item is to be given next. With objectively scoreable items, the procedure would be duck soup for a computer, but we felt that the complexities would probably sink a human examiner, especially if open-ended items were used. So reluctantly we abandoned the 'stradaptive' model and reverted to the two-stage model for our examiner-administered test. It is simpler to program for a computer than the complete Bayesian solution so that I expect to see this model quite common in computer-based adaptive tests in the future.

As you can see, Bayesian thinking is very like commonsense thinking. You start with your present best judgment, you seek the additional data that will add the most to that present judgment, and you put the old and the new together to arrive at a conclusion. The conclusion is always subject to refinement by further data if present information seems insufficient. What Bayesian procedures undertake to do, whether in educational research, or educational measurement, or test design, is to clarify the ground rules of this process so that information can be efficiently gathered and consistently and optimally synthesized so as to yield the least biased and most accurate result - no mean undertaking.